Datasheet

Yuxuan Wang

Applied data science

Data: Mar. 10th

Motivation for dataset creation

Why was the dataset created?

NBA player stats per game was created to record players' performance in 2017-2018 season such as average points, steals, turnovers per game. From the data, it quantitively shows the players' contribution. As a historic sports game, these data should be recorded for current analysis and future comparison.

What (other) tasks could the dataset be used for ?

It can be used to help players themselves and their coaches to find their flaws or advantages and customize their own training plan. Also, NBA fans can easily find their favorite players' data. Every year, several awards will be given to certain players and these data are important standard.

Has the dataset been used for any tasks already?

The basketball video game such as 2k and nbaonline use the data to create the ability number of player in the game.

Who funded the creation of the dataset?

The data was funded by The NBA reference. Com.

Dataset composition

What are the instances?

Each instance is the a player's average performance per game. Some players have multiple rows of data because they are traded to different teams.

Are relationships between instances made explicit in the data?

When players are traded to different team. He would have multiple rows of data. In the team column, TOT means the total average performance in all the team, which is the sum of the performance in each teams.

How many instance of each type are there?

There are totally 665 rows of players data. And there are total 540 players in the league 2017-2018 season.

What data does each instance consist of?

Each instance contains 3 txt columns which are players - name of the player, Pos - position in the game, Tm - player's team.
Other columns are all digital numbers such as games played in the season, points, per game, etc.

Is everything included or does the data rely on external resources?

The game video can be reviewed after each game. In most case, the stat per player is recorded during the game and will not be changed. The accuracy is high.

Are there recommended data splits or evaluation measures?

Usually, the data are accuracy. Sometimes, due to the referee, there might be some errors happened in fouls or turnovers, but we would use the official call as the reference.

What experiments were initially run on this dataset?

There is a formula to calculate a player's game efficiency.

Data Collection Process

How was the data collected?

The data is recorded while the game is processing. NBA official has its own data collection center.

Who was involved in the data collection process?

Unknown

Over what time-frame was the data collected?

After each game, the stat will be refreshed. Because my data is 2017-2018 season, the collect finished after the end of June 2018.

How was the data associated with each instance acquired?

Name is the player's name. Position is determined by their player profile though sometimes they are playing multiple positions. And the performance data is collected during the game.

Does the dataset contain all possible instance?

No, it can add factors such as salary.

If the dataset is a sample, then what is the population.

The population is all the basketball player's stat per game in the world. Containing other basketball association such as CBA (Chinese Basketball Association)

Is there information missing from the dataset and why?

Unknown.

Data Preprocessing

What preprocessing/cleaning was done?

The order of row is determined by their first letters of last name. Also, for the same player in different team, their Rk is the same.

Was the "raw" data saved in addition to the preprocessed/data?

No

Is the preprocessing software available?

Is, the function in lib pandas - read_csv can run the csv file.

Does this dataset collection/processing procedure achieve the mo- tivation for creating the dataset stated in the first section of this datasheet?

Yes, every player's stat is accuracy and is able to show their game performance. However, there is also limitations such as in different team, there are strong teams and weaker team. And for the all-star player, they would have more shooting chances.

Dataset Distribution

How is the dataset distributed?

The dataset can be download from NBA reference. Com in csv file. Also it can be copied to copy-boards and read by the pandas.

When will the datasets be released?

It is refreshed after each game. For the stat in 2017-2018 season, the final result would be released after the finals which is June, 2018.

What license (if any) is it distributed under?

The raw data belongs to NBA officials.

Dataset Maintenance

Who is supporting/hosting/maintaining the dataset?

The data is collected by the NBA officials and shared by NBA reference.com

Will the dataset be updated?

For this data of 2017-2018 season, it will not upgrade any more.

If the dataset becomes obsolete how will this be communicated?

In most cases, there will not be any change.

Is there a repository to link to any/all papers/systems that use this dataset?

There is a “share” button on this website that can transfer the table to several format options.

If others want to extend/augment/build on this dataset, is there a mechanism for them to do so?

Unknown.

Legal & Ethical Considerations

If the dataset relates to people (e.g., their attributes) or was gener- ated by people, were they informed about the data collection?

No, the data is got from public web site NBA reference

If it relates to other ethically protected subjects, have appropriate obligations been met?

Not applicable

If it relates to people, were there any ethical review applica- tions/reviews/ approvals?

It relates to people.

Unknown

If it relates to people, could this dataset expose people to harm or le- gal action?

There is minimal risk for harm: the data was already public.

If it relates to people, does it unfairly advantage or disadvantage a particular social group?

Unknown

Does the dataset comply with the EU General Data Protection Regu- lation (GDPR)?

The data does comply with GDPR.

Does the dataset contain information that might be considered sen- sitive or confidential?

No

Does the dataset contain information that might be considered inap- propriate or offensive?

No. It only contain name and numbers of stats.